

Critical Comparison

1 Introduction

XWRAP & STALKER allows you, to extract Data from HTML Websites.

Both of them need's a user interaction, to define the section of interests and rules, how to extract the right data.

2 Feature comparison

The XWRAP Wrapper is an four step based Wrapper. In step one the data is cleaned from syntactical incorrectness, in step two XWRAP parses the document, by each char algorithm, into an HTML-tree. In the next step the tree is parsed into a CSV file, and in the last step the user defined XML templates were used to generate a clean xml-output.

User input is necessary at step two, to define the section of interests. The XWRAP is not so intelligent to find out, which section has interest data, therefore the user has to define it.

STALKER uses a grammar called SLG to extract data out of documents – it is a learned based algorithm, which interact with the user. In this point XWRAP and STALKER has the same characteristic, no automation. The output of STALKER is not a document – it produces rules that are extraction content.

3. Performance results and example cases

All measurements of wrapper executions were carried out on a dedicated 200MHz Pentium machine(jambi.cse.ogi.edu). The machine runs Windows NT 4.0 Server and there is only one user in the system. All the XWRAP software is written in Java. The main Java package used is Swing.

XWRAP

Data Source	Generation Time(minutes)	Revision (times)	Extraction Rules Length(lines)	XML Template Length(lines)	Accuracy Verification
NCAA	40	2	114	153	100%
CIA Factbook	25	1	237	23	100%
Buy.com	16	0	102	46	100%
Stockmaster	23	1	90	46	100%

XWRAP Performance detail

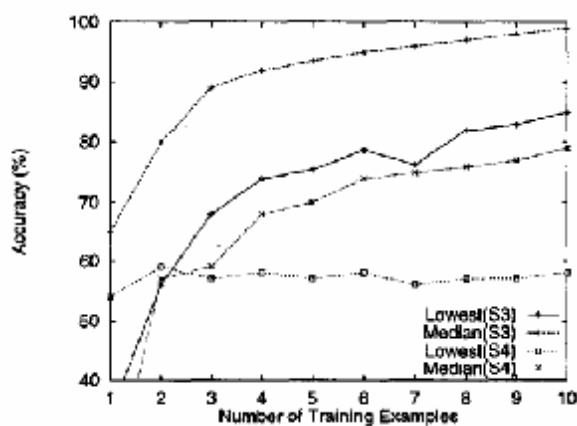
Data Source	Avg. vs. St. Dev.	Fetch Time(ms)	Expand Tree Times(ms)	Extraction Times(ms)	Generate Times(ms)	Total Time(ms)	Correlation Doc/Times
NCAA	Average	4391	8531	3841	1128	18520	0.45
	St. Dev.	1032	1055	228	116	1636	
CIA Factbook	Average	1907	11916	4709	3902	23043	0.93
	St. Dev.	265	3365	1175	1297	5776	
Buy.com	Average	6908	7777	2748	838	18909	0.66
	St. Dev.	4333	1553	1439	287	6602	
Stockmaster	Average	1972	5489	1412	468	9973	0.35
	St. Dev.	489	453	497	121	1131	

STALKER Experimental Data

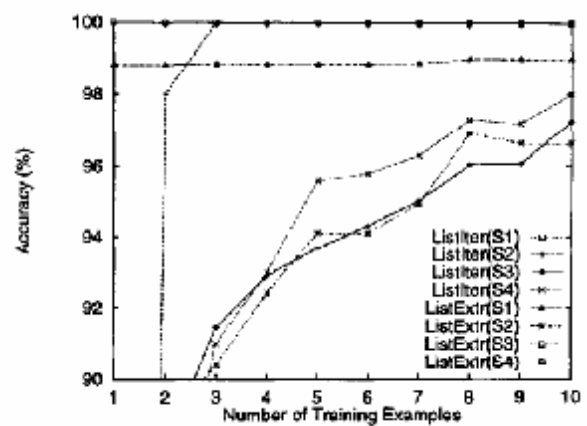
SR	Missing Items	Various Orders	WIEN		STALKER	
			Exs	CPU	Exs	CPU
S1	-	-	46	5 sec	1	19 sec
S2	-	-	274	83 sec	8	7 sec
S3	✓	✓	-	-	10	202 sec
S4	✓	-	-	-	10	708 sec

SRC	Extraction Task	Correctness	Nmb. Examples
S1	Score	100%	1
	Email	100%	1
	Name	100%	1
	FirstEntered	100%	1
	List Extraction	98.89%	10
	List Iteration	100%	1
S2	Name	100%	3
	Address	100%	3
	City	100%	4
	State	99.63%	10
	AreaCode	98.38%	10
	Phone	98.38%	10
	List Extraction	96.63%	10
	List Iteration	100%	3

STALKER learning details



(a) The median and lowest accuracy tasks for S3 and S4.



(b) List extraction and list iteration tasks.